

Sound Source Separation in a Multi-voice Environment Based on Auditory Central Nervous System

Yuan Luo¹, Kaiguo Tong¹, Yi Zhang¹, Huosheng Hu², Jun Chen¹, Chunjiang He¹

¹National Engineering Centre for Information Accessibility, Centre of Intelligent System & Robot
Chongqing University of Posts & Telecommunications, Chongqing, China

²School of Computer Science & Electronic Engineering, University of Essex, United Kingdom
luoyuan@cqupt.edu.cn; 359018647@qq.com; zhangyi99@263.net; hhu@essex.ac.uk

Abstract—Inspired by biology, this paper investigates voice processing method based on the central auditory system of the human brain. An integrated voice separation model is established by simulating central auditory system. Voice signals are processed under multi-spectral analysis by using peripheral auditory model. A coincidence neuron model is deployed to extract the features from voice signals so that the voice is separated in the cell model of brain inferior colliculus. Experimental results show that the integrated voice separation model can separate the voice in a multi-source environment reliably.

Keywords—Multi-source Environment; Auditory Central Nervous System; Intramural Time Difference; Binaural Level Difference; Voice Separation

I. INTRODUCTION

Up to now, multi-source speech recognition research has been conducted by worldwide researchers for many years, and has resulted in three popular approaches. The first approach is the application of three cues for sound localization, which was proposed by Rodemann et al. And it did not take advantage of the biological connections between the superior oliver complex (SOC) and inferior colliculus (IC) [1]. The second approach was proposed by Voutsas and Adamy, in which a multi delay-line model is developed by using spiking neural networks (SNN) incorporating realistic neuronal models. Their model only takes into account ITD and is not effective for frequencies higher than 1.5 kHz [2]. The third approach was proposed by Willert and Nixim, in which a probabilistic model is built to estimate the position of the sound sources, and the Bayesian theorem is used to calculate the connections between them. However, this method did not use spiking neural network to simulate realistic neuronal processing [3].

However, all the approaches mentioned above have good signal separation performance in a specific lab environment [4]. Their performance will be greatly degraded in a multi-source speech environment since the model does not completely simulate the central auditory system's role of separation in voice. During the past twenty-five years, our study for auditory central nervous system has made significant progress and shown that inferior colliculus (IC) plays a crucial role in the process of getting auditory information. IC is a hub and processing centre of extracting sound features [5]. Here, binaural time difference (ITD) and level difference (ILD) are both extracted from the sound.

Researches on hearing show that our two ears can identify features according to the differences of distance from source to each ear and the masking condition difference in the way of

3sound transmission. There are binaural time differences and binaural level difference when sound arrives from a place to each ear. As shown in Fig. 1, the time from sound source to each ear is different. When inputted speech information is separated by auditory nerve centre, the time difference and level difference are an important basis for sound separation [6].

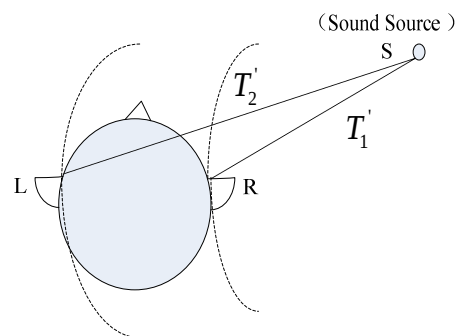


Fig.1 Binaural time difference (ITD) between two ears

IC will control the auditory cilia of inner ear nerve to response threshold, low frequency (<1.5kHz) voice signal will pass through the central area of the medial superior olive (MSO) to IC, (In this frequency range, ITD is more efficient to voice separation). While high frequency (>1.5kHz) voice signal will go through the central area of both medial superior olive (MSO) and Lateral superior olive (LSO) to IC, in this frequency range, ILD is more efficient to voice separation. At last, signals from different area enter IC separately [7].

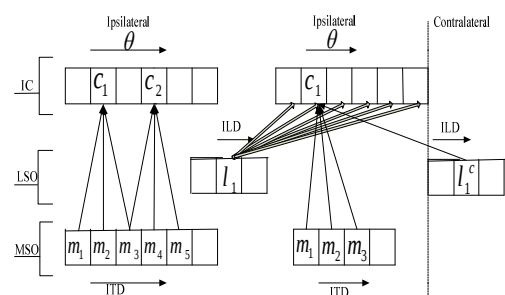


Fig.2 MSO and LSO analysis of the speech signal spectrum

Another important feature in the nerve tissue of IC is, physically, the use of a multi-layer anatomical structure to decompensate sound signals is in accordance with frequency. Each layer of never cells only responds to specific frequency

components. This anatomical feature is called frequency anatomical feature, which makes multi-band input audio isolated into space in the IC^[8]. Thus, the sounds from the same source or with the same frequency characteristics can easily be coincided and extracted. In the noisy multi-source environment, then speech signal is separated and extracted to re-generate signal flows^[9].

In summary, the auditory central nervous system can effectively separate input multi-source noisy, and a complete voice separation model which simulates the auditory central nervous system is possible to solve the voice recognition problem under a dynamic and complex environment.

II. THE SPEECH SEPARATION MODEL BASED ON CENTRAL AUDITORY SYSTEM

Fig. 3 is the schematic diagram of separation system based on auditory central nervous system, which is a complete calculation model of a simulating auditory central system. Multi-channel audio signals are firstly divided into different frequency channels according to different frequencies when going through the peripheral auditory model. Then voice information is extracted when through Oliver complex (SOC, including MSO and LSO). Finally, the multiple sound sources are separated into individual voice signals in the IC cells^[10].

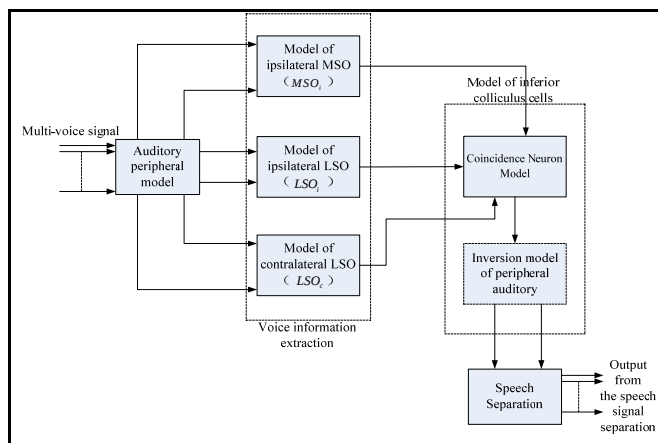


Fig.3 The speech separation model based on central auditory system

A. Auditory Peripheral Models

Acoustic studies have shown that the human ear canals have different responses to information with different frequencies. The basement membrane located inside the cochlea is an important link to the auditory central system, and can decompensate frequency. Different frequency signals will stimulate different vibrations in different locations of basement membrane^[11].

Based on these features of basement membrane, the processing of audio frequency peripheral uses the 16 second-order discrete Gammatone filters, which is used for multi-frequency analysis. The frequency options range from 100Hz to 4 kHz, which are chosen separately to do frequency decomposition for left and right ear signal aliasing by time frame. Each second-order discrete filter has a z-domain form,

$$F_i(z) = \left(\frac{Tz}{z^2 - 2e^{-b_i T} \cos(w_i T) + e^{-2b_i T}} \right)^4$$

$$\prod_{j,k=1}^2 = (z - e^{-b_i T} \cos(w_i T) + (-1)^k \sqrt{3 + (-1)^j 2^{1.5}} \sin(w_i T)) \quad (1)$$

Where i is the i -th number of filters in the group, w_i represents the frequency characteristic of each filter, b_i means the fixed bandwidth of each frequency.

Fig. 4 is a set of frequency response graphs of Gammatone filters with the auditory peripheral model:

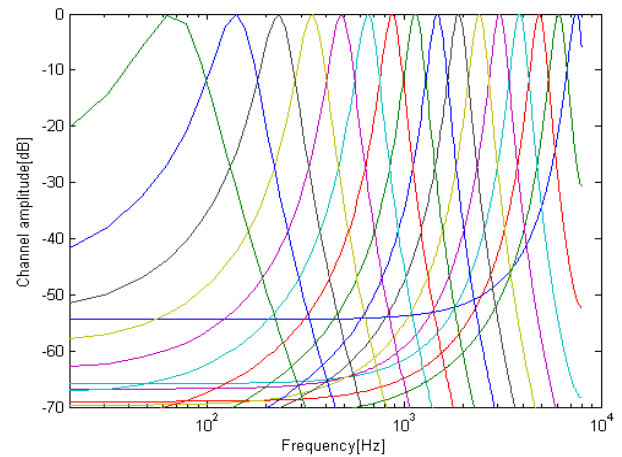


Fig. 4 The frequency response of filter group consisting of Gammatone filters

In Fig. 4, a filter group consists of 16 Gammatone filters with a range of frequency from 100 Hz to 4 kHz. For the input speech signal, after multi-frequency analysis of the auditory peripheral model, it passes through the auditory central system in 16 different frequency channels according to frequency difference.

B. Coincidence Neuron Model

Coincidence neuron model imitates the response of model of synaptic and neuronal to complete extraction and integration for audio information.

1) Generic Synaptic Model

Voice signal (wave) is spread by neurotransmitters when going through inner hair cell in the auditory central system. The vibration caused by wave on basement membrane will create neurotransmitters release through permeable membrane to synaptic cleft, and the movement of the hair brings the auditory nerve to a firing. The penetration of permeable membrane, $h(t)$, will change according to the input signal amplitude. Each Gammatone filter output will go through HWR.

$$h(t) = \begin{cases} gdt \left[\frac{x(t) + A}{x(t) + A + B} \right], & [x(t) + A] > 0, \\ 0, & [x(t) + A] \leq 0 \end{cases} \quad (2)$$

Where A is the percolation threshold produced by signal $x(t)$, g is a data related to penetration, dt means sampling interval, and B associates with the maximum penetration.

The model assumes that the inner hair cell could manufacture transmitter substance. The amount of transmitter substance in free store lying close to the membrane is $q(t)$, and it is restored at a rate $y[1-q(t)]$. The amount of transmitter substance in the synaptic cleft is $c(t)$, where the amount of being re-up-take into the hair cell is $rc(t)$. At the same time, some of it is lost from cleft and from the system altogether at a

rate $lc(t)$, which is summarized as follows.

$$\frac{dq}{dt} = y[1 - q(t)] + rc(t) - h(t)q(t) \quad (3)$$

$$\frac{dc}{dt} = h(t)q(t) - lc(t) - rc(t) \quad (4)$$

Then the auditory-never firing rate is:

$$p(t) = hc(t)dt \quad (5)$$

Equations (3), (4) and (5) consist of inner hair cell and synapse model. Where the g , y , r , l and h are constant and have the proper value, dt is the sampling interval.

2) Generic Soma Model

The transmitter molecules diffuse to the postsynaptic neuron through the synaptic cleft. The decay of the transmitter concentration is simulated by a leaky integrate-and-fire model. The amount of transmitter molecules on the postsynaptic neuron changes its permeability to certain ions. Ionic channels thus open gradually, receiving even more ions and forming a current. This procedure is modelled using a conductance mechanism which forms a gradually increasing postsynaptic current.

Leaky integrate-and-fire model that has been proposed as the model of neurons can be used for processing time-varying signals^[12] in powerful computing systems. The form of an integrate-and-fire model consists of a resistor R in parallel to a capacitor C which is charged by a current $I(t)$ in order to simulate the procedure of the postsynaptic current charging the ITD coincidence model. The voltage $U(t)$ across the capacitor C is compared to a threshold which is ϕ . The schematic diagram of the leaky integrate-and-fire model is shown in Fig. 5.

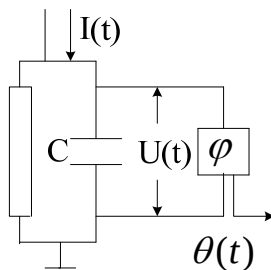


Fig. 5 Schematic of the leaky integrate-and-fire model

Between spikes, the voltage of a leaky integrate-and-fire model depends on:

$$U(t) = U_r \exp\left[-\frac{t-t_i}{RC}\right] + \frac{1}{C} \int_0^{t-t_i} \exp\left[-\frac{s}{RC}\right] I(t) ds \quad (6)$$

Where, U_r is the initial membrane potential, a typical value for RC is 1.6ms. When $U(t) = \phi$ at time t an output spike $\theta(t)$ is generated, and then $U(t)$ is reset to the initial voltage U_r .

3) Coincidence Neuron Model

In Figure 6, ITD pathway, the spike sequence of the contralateral ear passes variable delay line Δt_i . The delayed

spike sequence is denoted as $S_c(\Delta t_i, f_j)$, where C stands for the contralateral, Δt_i for the delay time, f_j for the frequency channel j . Similarly, $S_i(\Delta T, f_j)$ represents the delayed spike sequence of the ipsilateral ear with a fixed delay time ΔT . $S_c(\Delta t_i, f_j)$ and $S_i(\Delta T, f_j)$ are then input to the ITD coincidence model to calculate the ITD. The calculated output of the ITD coincidence model is a new spike sequence $S_{im}((\Delta T - \Delta t_i), f_j)$. If there are spikes in $S_{im}((\Delta T - \Delta t_i), f_j)$, it means the sound arriving to the ipsilateral ear is earlier than that to the contralateral ear by $ITD = \Delta T - \Delta t_i$ second. A negative ITD value means the sound arriving at the right ear is later than that at the left ear. As shown in Fig. 6, ES stands for excitatory synapse

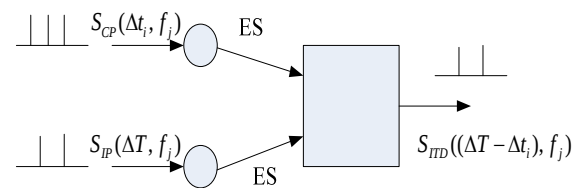


Fig. 6 ITD coincidence model

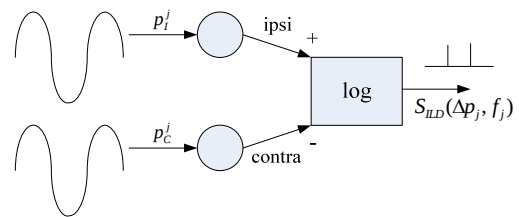


Fig. 7 ILD coincidence model

In Figure 7, the ILD pathway is modelled without using the leaky integrate-and-fire model. Instead, the detected sound levels of two sides are compared to calculate the level difference and the corresponding ILD cell is fired with a spike.

The level difference can be calculated as $\Delta p^j = \log\left(\frac{p_i^j}{p_c^j}\right)$,

where p_i^j and p_c^j stand for the ipsilateral and contralateral sound levels for the frequency channel j respectively. Similar to the ITD model, the output of the ILD model can be represented as a 2D spiking map, $S_{ild}(\Delta p_j, f_j)$. A negative ILD value means the level of sound going to the right ear is lower than that to the left ear. As shown in Fig. 7 ipsi and contra stand for ipsilateral and contralateral Gammatone frequency channels.

Before merging ITD and ILD maps for broadband sound separation, we must consider the advance of ITD and ILD in different frequency bands. On one side, ITD can accurately extract the features from voice signals up to 1.5 kHz but it begins to be ambiguous when frequency is above 1.5 kHz. On the other side, ILD is very trivial below 1.5 kHz. So we build two weight arrays, ITD_w and ILD_w , over frequency and compute either weighted ITD or ILD map by multiplying the weight array with the matrix of a 2D ITD/ILD map.

$$ITD_w^j = \frac{\sum_j (\max(\frac{f_j}{1500}, 1))}{\max(\frac{f_j}{1500}, 1)} \quad (7)$$

$$ILD_w^j = \frac{\max(\log(\frac{f_j}{1300}), 0)}{\sum_j (\max(\log(\frac{f_j}{1300}), 0))} \quad (8)$$

Here j is the channel index. Weighted ITD and ILD were eventually integrated into the mapping information together, which is the output of MSO and LSO, and finally entered into the hypothalamus of the brain nerve cells to the voice information extraction and separation.

C. IC Cell Model

The cells in the IC can be classified into 6 physiological types including Sustained-regular, Rebound-regular, Onset, and so on. We tested the Onset cell which is among the 6 types. The Onset cell of IC is simulated in order to separate the speech signal in this paper. Fig. 8 is the structure of Onset cell of IC.

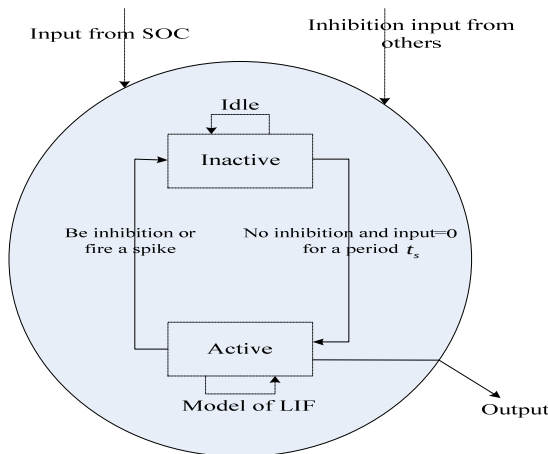


Fig. 8 The IC's onset cell

In general, Onset cell has two states: active and inactive. In Fig.8, when the cell is active, the cell is implemented as a LIF neuron until a spike is fired, or the cell receives an inhibitory input after which the cell state becomes inactive. The cell goes back to active state only when there is no inhibition and the input is 0 (no spike) for a period of t_s .

Onset cell model is re-used for multi-source speech signal separation. First, the speech signal energy in the nerve cell model of the i -th frequency channel and the j -th time frame $\sum_i S_{i,j(t)}^2$ and the noise signal energy $\sum_i S_{i,j(t)}^2$ and energy $\sum_i n_{i,j(t)}^2$ is calculated. Then the signal energy ratio is calculated as follows.

$$E_{i,j} = \frac{\sum_i S_{i,j(t)}^2}{\sum_i S_{i,j(t)}^2 + \sum_i n_{i,j(t)}^2} \quad (9)$$

On one hand, when $E_{i,j} > 0.5$, it means if the voice energy is greater than the noise energy, this voice should be remained.

On the other hand, when $E_{i,j} < 0.5$, this voice should be discarded. The Onset cell model was then re-used to obtain the ITD and ILD values to build a masking matrix for separating the voice signal. Here, binary masking for the i -th frequency channel and the j -th time frame of the masking coefficient can be defined as,

$$\lambda(i, j) = \begin{cases} 1 & \text{if } [(f_i \leq f_c) \text{ and } [\tau_{\max}(i, j)] > T^{(r)}(i, j)] \\ & \text{if } [(f_i > f_c) \text{ and } [L(i, j)] > T^{(l)}(i, j)] \\ 0 & \text{else} \end{cases} \quad (10)$$

Where $f_c = 1.5$ kHz, $T^{(r)}(i, j)$ and $T^{(l)}(i, j)$ are the threshold separately for ITD and ILD. $\tau_{\max}(i, j)$ is the max delay of the i -th frequency channel and the j -th time frame. $L(i, j)$ is the value of ILD of the i -th frequency channel and the j -th time frame.

$$L(i, j) = 20 \lg \frac{\sum_{i,j} p_l(i, j, t)^2}{\sum_{i,j} p_r(i, j, t)^2} \quad (11)$$

Where $P_l(i, j, t)$ and $P_r(i, j, t)$ are the signal firing rate of left and right ears of the i -th channel and the j -th time frame.

Multi-source speech signal in each frequency channel and time frame of the masking coefficients are calculated and then get masking matrix. Likewise matrix elements 1 and 0 are used respectively for the same ownership. All the elements 1 are in the same matrix. Actually, the Fourier transform of its autocorrelation is equal to the square of its Fourier transform magnitude. It is assumed that $R_{xx}(\tau)$ is the autocorrelation of $x(t)$. Then the power spectrum of $x(t)$ is,

$$|X(w)|^2 = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{-jw\tau} d\tau \quad (12)$$

Therefore, by simple operations, we can obtain the magnitude of the short time Fourier transforms. The remaining problem is how to reconstruct signals only from the magnitude of the Short Time Fourier Transform(STFT). To achieve this operation, the iterative algorithm is used to reconstruct the phase of the signal at each of iteration in order to decrease the square error between the STFT magnitude of the reconstructed signal and the STFT magnitude which is known already. Then, we can obtain the estimate value and minimize the square error between the magnitude of the STFT of the signal reconstructed and the magnitude of the STFT which is known already.

The reconstructed signal by i iterations is,

$$x^{(i)}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n) \frac{1}{2\pi} \int_{-\pi}^{\pi} X^{(i-1)}(m, n) e^{jw\tau} dw}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} \quad (13)$$

Where $w(mS-n)$ is the analysis window and S is the window shift.

Then we can obtain the STFT $X^{(i)}(m, n)$ of the i -th reconstructed signal in terms of $x^{(i)}(n)$, and acquire the error between which the primary given short-time magnitude use equation as follows,

$$Error = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{N-1} \|X^{(i)}(m, n) - |X_d(m, n)|\|^2 \quad (14)$$

If the error is smaller than the given value, iteration is over, or what we should do is to calculate $\hat{X}^{(i)}(m, n)$, and run the $i+1=th$ iteration. $\hat{X}^{(i)}(m, n)$ is as follow.

$$\hat{X}^{(i)}(m, n) = |X_d(m, n)| \frac{X^i(m, n)}{|X^i(m, n)|} \quad (15)$$

Thus the auditory-nerve firing rate $p(t)$ of each channel of the auditory model could be calculated by correlogram. The next step is to reconvert the signal $h(t)$ coming out of the HWR stage in terms of $p(t)$.

$$c(t) = \frac{p(t)}{hdt} \quad (16)$$

Then $q(t)$ and $h(t)$ could be obtained as follows.

$$q(t) = y[1 - q(t-1)]dt - lc(t-1)dt - c(t) - c(t-1) + q(t-1) \quad (17)$$

$$h(t) = \left[\frac{c(t) - c(t-1)}{dt + lc(t) + r(t)} \right] / q(t) \quad (18)$$

Where $h(t)$ is the signal expression coming out of the HWR stage.

Following the reverse process of the HWR is to reconstruct the negative parts of the signal after re-iterating.

III. EXPERIMENTAL RESULTS

To verify the correctness and effectiveness of the above model, many experiments are implemented. In the experiments, Onset cell is investigated on a PC with the Intel Pentium 2.5GHz and 1GBits memory, and Matlab-Simulink is used in the tests at a sampling rate 44.1 kHz and 16 bit sampling precision. The three types test data of all are summarized in this paper, which are class A, class B and class C. Each set of data includes voice signals and noise signals (or other interference voice) respectively.

Class A includes sources which are composed by Chinese words, and source two composed by traffic noise. Class B includes sources which are composed by Chinese phrases, and source two composed by the traffic noise. Class C includes source which composed by Chinese female word, and source two composed by Chinese male word.

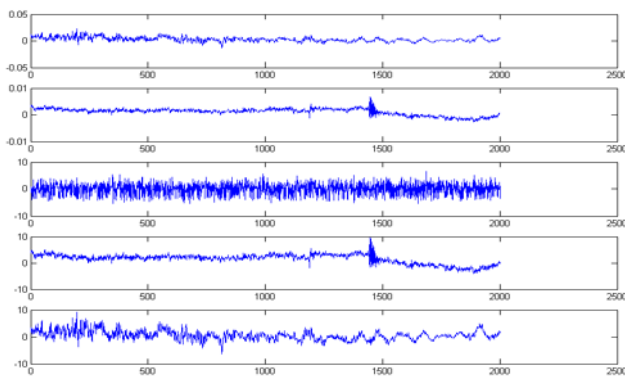


Fig. 9 The signal waveforms of simulation results in the third group

Figure 9 is the signal waveform comparison simulation results of group C. The first line of signal waveform is the original female voice "forward". The second line is the original male voice "forward". The third one is the signal after aliasing. The fourth line is the separation of the source signal with the male voice "forward". The fifth one is the separation of source signals with female voice "forward".

For the tests of three types, class A, class B and class C, a large number of repetitive tests have been implemented. In this paper, 20 groups test results of each class were randomly selected, and the separated voice signal and the original speech waveform will be compared by Matlab.

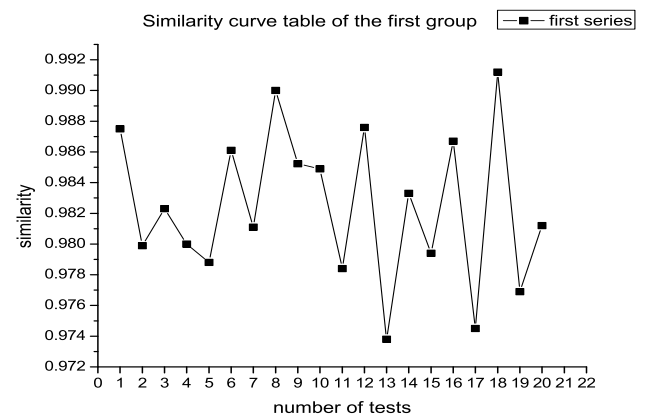


Figure 10. Similarity curve table of the first group

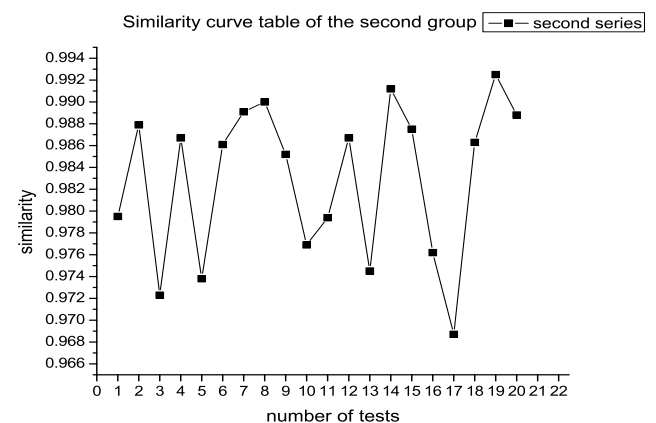


Fig. 11 Similarity curve table of the second group

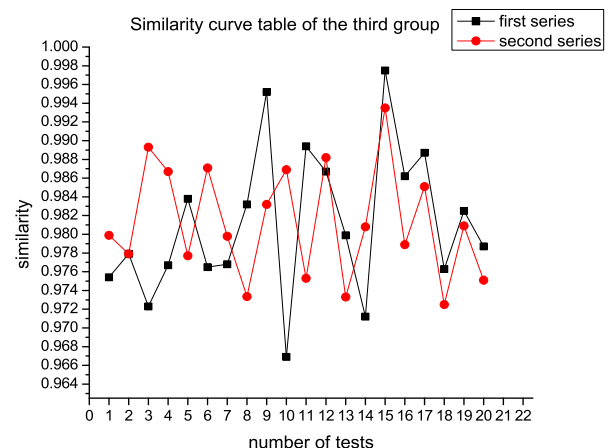


Fig. 12 Similarity curve table of the third group

Fig. 10, Fig. 11 and Fig. 12 present the results of similarity comparison in three groups. Abscissa represents the number of tests; the vertical axis represents the similarity of the separated and original voice signal. As shown in the graph, the similarities of the separated voice signal and the original signal are more than 0.97. As a result, the robustness of the method for the separation of speech signals is high.

Recently, Voutsas and Adamy built a multi delay-line model using spiking neural networks incorporating realistic neuronal models. Their model only takes into account ITD. The average similarity reaches 0.975 when the frequency of sound is less than 1.5 kHz, but it is not effective for the frequency which is higher than 1.5 KHz. Fig. 13 presents the results of similarity comparison of the model of Voutsas and Adamy. Test data from 1 to 25 show the model of voutsas and Adamy used in the frequency which is less than 1.5 kHz. Test data from 25 to 50 shows the model of voutsas and Adamy used in the frequency which is more than 1.5 kHz.

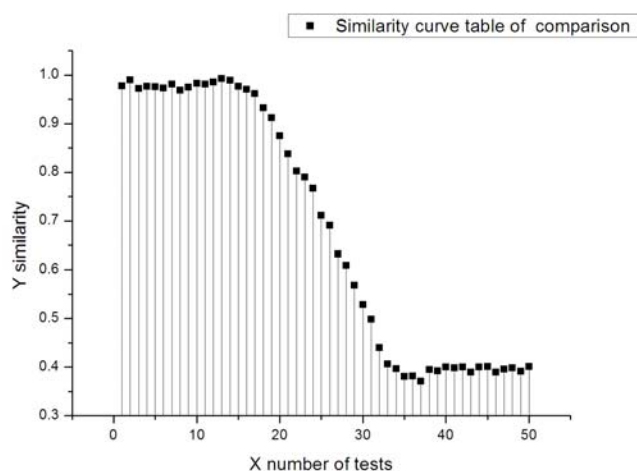


Fig. 13 Similarity curve table of the model of voutsas and Adamy

From Fig. 13, we can see that the model of Voutsas and Adamy is only used for the sound frequency less than 1.5 kHz. But the model proposed in this paper is used for the sound frequency higher than the frequency ranges.

For experiments of class A and class B, using this approach can improve signal- to- noise ratio of the speech. This coincidence neurons model in the integration of the information ITD and ILD selects five of the azimuth data. Signal- to- noise ratio is calculated according to:

$$SNR(dB) = 10 \lg \left(\frac{\sum_i \hat{s}(t)^2}{\sum_i [s(t) - \hat{s}(t)]^2} \right) \quad (19)$$

Table 1 shows the results from comparing, in which “Before” means “Before separation” and “After” means “After separation”.

TABLE 1 THE CONTRAST OF SIGNAL TO NOISE RATIO

Azimuth	First group		Second group	
	Before	After	Before	After
0 , 30	15.8	50.2	13.7	48.2
0 , 60	14.6	49.7	12.9	47.8
45 , 90	14.5	49.1	12.8	48.1

100, 120	14.7	49.6	13.1	47.9
135, 140	12.8	20.7	11.7	21.6

As we can see: when two sound sources are coincided with certain spatial orientation differences, the separation of the signal- to- noise ratio has been greatly improved. When the incident spatial orientation difference of two sound sources is small, the separation of the voice signal- to- noise ratio is not very different to one that is before separation. For example, when the azimuth (θ_1, θ_2) is selected as (135,140), the coincidence neurons model in the calculation information of ITD and ILD is likely to cause bias, then resulting in masking coefficient calculation error. This phenomenon can also be used to explain the phenomenon of human hearing. For example, when the two sounds come from very similar azimuth, the human auditory system is difficult to distinguish them.

IV. CONCLUSION AND FUTURE WORK

This paper proposes a sound separation method in the environment with multi-voice sound sources. A model of the human brain auditory central nervous system was completely established. Compared with the existing voice recognition, this model is a good solution to the problems that a vast majority of speech recognition methods can only be used in an environment with single sound source and low-noise.

In the further research, the central auditory system model can be firstly applied to the speech recognition of mobile robots to achieve a high voice recognition rate; Secondly, this model can be used in hearing aids for hearing-impaired people; Thirdly, this model can assist the current text retrieval; Fourthly, the biological model can be used in sound field reconstruction and virtual- reality environment.

ACKNOWLEDGMENTS

This research is financially supported by International Science and Technology Cooperation Program of China, 2010DFA12160, National Natural Science Foundation of China, 51075420, and Science& Technology Research Project of Chongqing CSTC, 2010AA2055.

REFERENCES

- [1] T. Rodemann, M. Heckmann, F. Joubin, C. Goerick, and B. Scholling. Real-time Sound Localization With a Binaural Head-system Using a Biologically-inspired Cue-triple Mapping. Proc. of 2006 IEEE/RSJ International Conference on Intelligent Robots & Systems. 2006, pp. 860-865.
- [2] K. Voutsas and J. Adamy. A biologically inspired spiking neural network for sound source lateralization. IEEE Trans. on Neural Networks, ISSN 10459227.vol. 18(6), 2007:pp. 1785-1799.
- [3] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. Journal of the Acoustical Society of America, vol. 119, 2006:pp. 463-479.
- [4] D. Oertel, R.R. Fay, and A.N. Popper, Integrative Functions in the Mammalian Auditory Pathway. 2002, Springer: New York. 431.
- [5] S. Wermter, et al., Towards multimodal robot learning. Robotics and Autonomous Systems Journal, 2004.Vol. 47, No. 2-3, pp. 171-175.
- [6] H.M. Zhao, L. Geg, X.Q. Chen. Research based on sound localization and auditory masking effect of voice separation[J]. Electronics, 2005, 1(1):158-160.
- [7] J.D. Liu, H. Erwin and S. Wermter. Mobile Robot Broadband Sound Localisation Using a Biologically Inspired Spiking Neural Network. Proceedings of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems in Nice. 2008. pp. 2191-2196.
- [8] T. Rodemann, M. Heckmann, F. Joubin, C. Goerick, and B. Scholling. Real-time Sound Localization With a Binaural

- Head-system Using a Biologically-inspired Cue-triple Mapping. In *Intelligent Robots and Systems*, 2006 IEEE/RSJ International Conference on. 2006, pp. 860-865.
- [9] C. Panthey and S. Wermter. Spike-timing-dependent synaptic plasticity: from single spikes to spike trains. *Neurocomputing*, 2004, Vol 58-60, pp. 365-371.
- [10] H.M. Zhao, Y.Q. Wang, X.Q. Chen. Auditory model inversion and its application[A]. *Proc. ICNNSP' 03[C]*. Nanjin, China: ICNNSP, 2003, 868-878.
- [11] M. Slaney, "An efficient implementation of the Patterson -Holds worth auditory filter bank," *Apple Computer*, Cupertino, CA, Tech. Rep.TR-35, 1993
- [12] W. Gerstner, "Integrate-and-Fire Neurons and Networks," in the *Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., 2nd ed. Cambridge, MA: MIT Press, 2003, pp. 577-581.

AUTHORS BIOGRAPHY

Yuan Luo is a Professor in School of Optical Engineering, Chongqing University of Post and Telecommunications. She obtained her PhD degree from Chongqing University in 2003. Her research interests include signal and information processing and digital image processing.

Kaiguo Tong is currently an MSc student in the Centre for National Engineering Research & Development for Information Accessibility at Chongqing University of Post and Telecommunications. His research interests focus on speech

recognition.

Yi Zhang was born in Chongqing, in 1966. He received PhD from Huazhong University of Science and Technology in 2002, and received post doctorate from Southeast University in 2005. Now he is a professor in Chongqing University of Posts and Telecommunications. His research interests mainly include robots and applications, data fusion.

Huosheng Hu is a Professor in School of Computer Science & Electronic Engineering at the University of Essex, leading the Human-Centred Robotics Group. His research interests include behaviour-based robotics, human-robot interaction, embedded systems, mechatronics, learning algorithms, and pervasive computing. He has published over 300 papers in journals, books and conferences in these areas, and received a number of best paper awards. He is one of founding members of IEEE Robotics & Automation Society Technical committee on Networked Robots, a Fellow of IET and InstMC, a senior member of IEEE and ACM. He has been a chair or committee member for many international conferences such as IEEE ICRA, IEEE IROS, IEEE ROBIO, IEEE ICMA and IEEE ICIA conferences. He is currently Editor-in-Chief for International Journal of Automation and Computing.